混合高斯分布和EM算法

Wu X.J.

2020年4月23日

目录

1 高斯混合模型

1.1 数学模型

定义一维高斯分布（或称为正态分布）为$x \sim \mathcal{N}(\mu, \sigma^2)$，其概率密度函数是

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

图1显示了一维正态分布的概率密度函数的情况。

而对于多维的高斯分布，考虑n个随机变量组成的随机变量向量$\boldsymbol{x} \in \mathbb{R}^n$，期望为$\boldsymbol{\mu}$，其中$\mu_i$表示随机变量$x_i$的期望，以及协方差矩阵$\boldsymbol{\Sigma} \in \mathcal{M}_{n\times n}$，于是多维随机变量的概率密度函数是

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

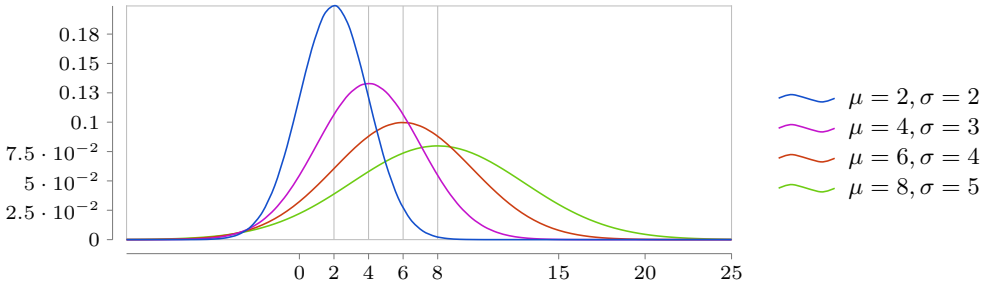

图 1: 一维正态分布中不同的期望和方差的选择

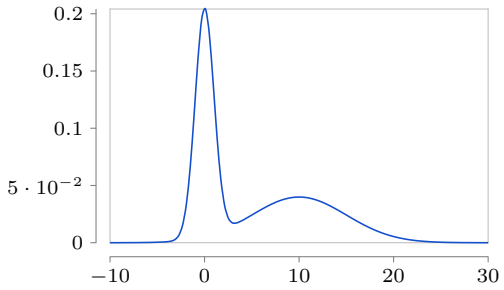

图 2: $x \sim \mathcal{N}(0,1)$和$x \sim \mathcal{N}(10,5)$按照权重系数$0.5, 0.5$混合后的概率密度图

如果按照一定的权重将若干个高斯模型进行叠加，我们就得到了混合的高斯模型。具体来说，设$\alpha_i > 0 (i = 1, 2, \ldots, k)$且$\sum_i \alpha_i = 1$，则定义混合的概率密度函数为

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i f_i(\boldsymbol{x}) \tag{1}$$

其中

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T {\boldsymbol{\Sigma}_i}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right) \tag{2}$$

这仍然是满足概率密度的定义，因为

$$\int_{\boldsymbol{x}} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{x}} \sum_{i=1}^{k} \alpha_i f_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \int_{\boldsymbol{x}} f_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i = 1 \tag{3}$$

图2是一维情况下，将两个高斯模型$x \sim \mathcal{N}(0,1)$和$x \sim \mathcal{N}(10,5)$按照权重系数$0.5, 0.5$混合后的概率密度图像。

1.2 为什么需要混合高斯模型

设有若干数据的数据集，并且假定每一个数据由且仅由某一个特定的高斯分布所产生，如果我们使用单一的高斯模型去拟合此数据集（如极大似然估计）可能会得出不合理的模型出来。考虑图3的数据分布情况，其数

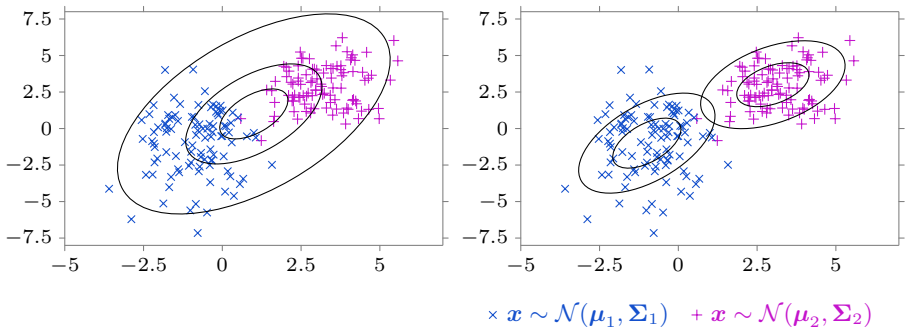

图 3：左：使用单一的高斯分布模型拟合数据集；右：使用混合的高斯模型来拟合。

据是由两个高斯分布所组成的。左图使用了单一的高斯分布模型去拟合，其结果显然不合理，因为在中心处分布的数据点反而少；右图使用了混合的高斯模型去拟合，其结果与实际情况比较符合。

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),\ \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

其中

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 5 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 2 \end{bmatrix}$$

混合高斯模型的本质就是融合几个单一高斯模型使得模型更加复杂，从而产生更复杂的样本。理论上，如果某个混合高斯模型融合的高斯模型个数足够多，它们之间的权重设定得足够合理，这个混合模型可以拟合任意分布的样本。

1.3 权重系数的解释

因为高斯混合模型的概率密度函数1中的系数a_i满足$0 \leqslant a_i \leqslant 1$，我们可以将其视为一个概率：即 **选择第$i$个模型的先验概率**。换句话说，给定一个随机变量$\boldsymbol{x}$，它有$a_i$的概率是由第$i$个模型所产生的，在此条件下$\boldsymbol{x}$的概率就是$f_i(\boldsymbol{x})$，因此可以将概率写为

$$\mathbb{P}(\boldsymbol{x}) = \sum_{i=1}^{k} \mathbb{P}_{model}(k)\mathbb{P}(\boldsymbol{x}|model = k) = \sum_{i=1}^{k} \alpha_i f_i(\boldsymbol{x}) \tag{4}$$

我们将在第2节中使用到这一思路。

2 GMM的极大似然估计

设有n个样本的样本集$Y = \{\boldsymbol{y}^{(i)}, i = 1, 2, \ldots, n\}$，每一个样本是一

个m维随机变量的观测值：$\boldsymbol{y}^{(i)} \in \mathbb{R}^m$。假设这些样本每一个观测值都**由且仅由$k$个高斯分布模型中的一个**所产生：$\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2, \ldots, k$，令

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}, \boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k\}$$

我们现在的任务是，通过极大似然估计来估计出参数$\boldsymbol{\mu}, \boldsymbol{\Sigma}$。写出极大似然函数

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{y}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5}$$

然后最大化L即可：

$$\boldsymbol{\mu}, \boldsymbol{\Sigma} = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6}$$

但是立刻就会有一个问题：我们并不知道任何一个观测数据$\boldsymbol{y}^{(i)}$到底是由哪一个高斯模型产生的，换句话说，样本来源这一数据丢失了。为此，我们假设样本来自第i个高斯模型的先验概率为$\mathbb{P}_{model}(i) = \alpha_i$，从而我们将样本概率写为

$$\mathbb{P}(\boldsymbol{y}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^{k} \alpha_j \mathbb{P}(\boldsymbol{y}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{7}$$

于是极大似然函数就是

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \alpha_j \mathbb{P}(\boldsymbol{y}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{8}$$

两边取对数，得到

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \alpha_j \mathbb{P}(\boldsymbol{y}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right) \tag{9}$$

然后只要令$\frac{\partial}{\partial \boldsymbol{\mu}} L = 0$以及$\frac{\partial}{\partial \boldsymbol{\Sigma}} L = 0$即可。但是仔细分析就会发现对数似然函数里面，对数里面还有求和。实际上没有办法通过求导的方法来求这个对数似然函数的最大值。此时就要用到EM算法。

3 EM算法

最大期望算法（Expectation-Maximization algorithm, EM），或Dempster-Laird-Rubin算法，是一类通过迭代进行极大似然估计的优化算法，通常作为牛顿迭代法的替代用于对包含**隐变量或缺失数据**的概率模型进行参数估计。EM算法的两个步骤：E-step(expectation-step，期望步)和M-step(Maximization-step,最大化步)。

3.1 E-Step

为每一个样本$\boldsymbol{y}^{(i)}$引入一个k维的隐变量(latent variable)$\boldsymbol{\gamma}^{(i)} \in \mathbb{R}^k$，如果该样本来自第$r$个高斯模型，那么$\gamma_r^{(i)} = 1$，其余分量均为0。换句话说，该隐变量指明了该样本是来自哪一个高斯分布模型。这样，$(\boldsymbol{y}^{(i)}, \boldsymbol{\gamma}^{(i)})$就构成了第$i$个样本的完整描述。

设先验概率$\Phi = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$，先考虑概率

$$\mathbb{P}\left(\boldsymbol{y}^{(i)}, \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right)$$

可以写成条件形式

$$\mathbb{P}\left(\boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) \mathbb{P}\left(\boldsymbol{y}^{(i)} | \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right)$$

根据隐变量的定义，样本$\boldsymbol{y}^{(i)}$来自于第r个高斯模型当且仅当$\boldsymbol{\gamma}^{(i)}$的第$r$个分量为1、其余分量为0，从而根据先验概率有

$$\mathbb{P}\left(\boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) = \alpha_r$$

另一方面，给定了$\boldsymbol{\gamma}^{(i)}$后我们就知道了这个样本是来自于哪一个高斯模型，假设是第r个，于是就有

$$\mathbb{P}\left(\boldsymbol{y}^{(i)} | \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) = f_r(\boldsymbol{y}^{(i)})$$

综上我们就有

$$\mathbb{P}\left(\boldsymbol{y}^{(i)}, \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) = \alpha_r \cdot f_r(\boldsymbol{y}^{(i)}) \quad \text{where } \gamma_r^{(i)} = 1$$

考虑到隐变量的取值特性，我们可以写成一个更加紧凑的形式

$$\mathbb{P}\left(\boldsymbol{y}^{(i)}, \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) = \prod_{j=1}^{k} \left(\alpha_j \cdot f_j(\boldsymbol{y}^{(i)})\right)^{\gamma_j^{(i)}} \tag{10}$$

于是我们给出基于隐变量（包括先验概率）的极大似然函数

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi) =& \prod_{i=1}^{n} \mathbb{P}\left(\boldsymbol{y}^{(i)}, \boldsymbol{\gamma}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi\right) \\ =& \prod_{i=1}^{n} \prod_{j=1}^{k} \left(\alpha_j \cdot f_j(\boldsymbol{y}^{(i)})\right)^{\gamma_j^{(i)}} = \prod_{j=1}^{k} \prod_{i=1}^{n} \alpha_j^{\gamma_j^{(i)}} f_j(\boldsymbol{y}^{(i)})^{\gamma_j^{(i)}} \\ =& \prod_{j=1}^{k} \left(\alpha_j^{\sum_{i=1}^{n} \gamma_j^{(i)}} \prod_{i=1}^{n} f_j(\boldsymbol{y}^{(i)})^{\gamma_j^{(i)}} \right) \end{aligned} \tag{11}$$

两边取对数，得到

$$\begin{aligned}\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma},\Phi) &= \sum_{j=1}^{k}\ln\left(\alpha_j^{\sum_{i=1}^{n}\gamma_j^{(i)}}\prod_{i=1}^{n}f_j(\boldsymbol{y}^{(i)})^{\gamma_j^{(i)}}\right)\\ &= \sum_{j=1}^{k}\left(\ln\alpha_j\cdot\sum_{i=1}^{n}\gamma_j^{(i)}+\sum_{i=1}^{n}\gamma_j^{(i)}\ln f_j(\boldsymbol{y}^{(i)})\right)\end{aligned}\tag{12}$$

接下来只要计算

$$\arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma},\Phi}\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma},\Phi)$$

但是到目前为止我们还不能对该对数极大似然函数最大化，因为隐变量$\boldsymbol{\gamma}$确实还是不知道。因此这里的思路就是使用迭代：

> 我们初始时可以给出参数$\boldsymbol{\mu},\boldsymbol{\Sigma},\Phi$的一个估计值$\boldsymbol{\mu}^{(0)},\boldsymbol{\Sigma}^{(0)},\Phi^{(0)}$（通过随机初始化或者是从样本中估计），通过$w$轮迭代优化后得到$\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)}$。我们使用本轮得到的参数对隐变量$\boldsymbol{\gamma}$做进一步的估计，然后用估计出的结果再进行迭代得到$\boldsymbol{\mu}^{(w+1)},\boldsymbol{\Sigma}^{(w+1)},\Phi^{(w+1)}$，以此类推。

具体来说，给定参数估计$\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)}$，对于给定样本点$\boldsymbol{y}^{(i)}$的条件下，我们首先计算其对应的隐变量中第$j$个分量的期望（因为分量只能取$0,1$，从而期望在数值上就等于该样本点来自于第$j$个高斯模型的概率）。计算期望的作用在于 **我们将使用$\boldsymbol{\gamma}$的期望来代替当前$\boldsymbol{\gamma}$的值，来进行下一轮迭代**[1]：

$$\begin{aligned}\mathbb{E}(\gamma_j^{(i)}|\boldsymbol{y}^{(i)};\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)}) =& \mathbb{P}(\gamma_j^{(i)}=1|\boldsymbol{y}^{(i)};\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)})\\ =& \frac{\mathbb{P}(\gamma_j^{(i)}=1,\boldsymbol{y}^{(i)};\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)})}{\mathbb{P}(\boldsymbol{y}^{(i)};\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)})}\end{aligned}$$

分母部分代表的含义是给定一个混合模型$\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)}$情况下，一个样本点$\boldsymbol{y}^{(i)}$的概率，按照我们之前的计算，它应该是

$$\mathbb{P}(\boldsymbol{y}^{(i)};\boldsymbol{\mu}^{(w)},\boldsymbol{\Sigma}^{(w)},\Phi^{(w)}) = \sum_{j=1}^{k}\alpha_j^{(w)}f_j^{(w)}(\boldsymbol{y}^{(i)})$$

[1] 另一种理解是考虑离散情况，多维随机变量函数$f(\boldsymbol{x})$中对某一个随机变量x_i求期望，写为

$$\mathbb{E}_{x_i}(f(\boldsymbol{x})) = \sum_{\xi\in x_i}p(\xi)\cdot f(\boldsymbol{x}|_{x_i=\xi})$$

于是期望可以理解为f，如果将f看作是x_i的函数的话，取值的平均情况。因此，在后文中我们会计算$\mathbb{E}_\gamma(\ln L)$，此时就是在求当γ取遍不同值时，极大似然函数函数$\ln L$的平均取值情况。

换句话说，因为γ指示了样本来源于哪一个模型，不同的γ就代表了样本点来自不同高斯模型的可能性，期望就是考虑到所有来源可能后，这些样本点的概率——也就是极大似然函数的意义——的平均值，此时如果我们计算出当参数为$\boldsymbol{\mu}',\boldsymbol{\Sigma}',\Phi'$时这个期望（平均值）达到最大，因为平均值提高了，就说明无论这些点来自于哪一个高斯模型（即给定某一γ），新的模型$\boldsymbol{\mu}',\boldsymbol{\Sigma}',\Phi'$比原来的模型$\boldsymbol{\mu},\boldsymbol{\Sigma},\Phi$都更有可能产生这些样本点，这就是极大化期望的意义。

其中$\alpha_j^{(w)} \in \Phi^{(w)}$，$f_j^{(w)}(\boldsymbol{y}^{(i)}) \sim \mathcal{N}(\mu_j^{(w)}, \Sigma_j^{(w)})$，均是根据第$w$次迭代的结果来计算的。同理，对于分子部分，表示的含义是给定混合模型$\boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}$情况下样本点$\boldsymbol{y}^{(i)}$确实来自于第$j$个高斯模型的概率，它等于选择该高斯模型的先验概率和样本点在此高斯模型中的概率的乘积，即

$$\mathbb{P}(\gamma_j^{(i)} = 1, \boldsymbol{y}^{(i)}; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}) = \alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})$$

于是我们就得到了所求的期望

$$\mathbb{E}(\gamma_j^{(i)} | \boldsymbol{y}^{(i)}; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}) = \frac{\alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})}{\sum_{j=1}^{k} \alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})} \tag{13}$$

我们对式(12)求关于隐变量γ的期望，记

$$\mathbb{E}_{\gamma_j^{(i)}}^{w} = \mathbb{E}(\gamma_j^{(i)} | \boldsymbol{y}^{(i)}; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)})$$

从而令

$$\begin{aligned} Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}) =& \mathbb{E}_{\gamma} \left(\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi); \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)} \right) \\ =& \sum_{j=1}^{k} \left(\ln \alpha_j \cdot \sum_{i=1}^{n} \mathbb{E}_{\gamma_j^{(i)}}^{w} + \sum_{i=1}^{n} \mathbb{E}_{\gamma_j^{(i)}}^{w} \ln f_j(\boldsymbol{y}^{(i)}) \right) \end{aligned} \tag{14}$$

这样我们只要（再）对Q做极大似然估计就可以得到下一轮的模型参数估计

$$\boldsymbol{\mu}^{(w+1)}, \boldsymbol{\Sigma}^{(w+1)}, \Phi^{(w+1)} = \arg \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi} Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}) \tag{15}$$

3.2 M-step

这一步就是对式(14)进行最大化

$$\begin{aligned} &\boldsymbol{\mu}^{(w+1)}, \boldsymbol{\Sigma}^{(w+1)}, \Phi^{(w+1)} \\ &\quad = \arg \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi} \sum_{j=1}^{k} \left(\ln \alpha_j \cdot \sum_{i=1}^{n} \mathbb{E}_{\gamma_j^{(i)}}^{w} + \sum_{i=1}^{n} \mathbb{E}_{\gamma_j^{(i)}}^{w} \ln f_j(\boldsymbol{y}^{(i)}) \right) \end{aligned} \tag{16}$$

其中

$$\mathbb{E}^{w}_{\gamma_j^{(i)}} = \frac{\alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})}{\sum_{t=1}^{k} \alpha_t^{(w)} f_t^{(w)}(\boldsymbol{y}^{(i)})} \tag{17}$$

$$f_j^{(w)}(\boldsymbol{y}^{(i)}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_j^{(w)}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)})^T {\boldsymbol{\Sigma}_j^{(w)}}^{-1} (\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)})\right) \tag{18}$$

$$f_j(\boldsymbol{y}^{(i)}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j)^T {\boldsymbol{\Sigma}_j}^{-1} (\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j)\right) \tag{19}$$

$$\sum_{j=1}^{k} \alpha_j = 1 \tag{20}$$

我们对方程求导并令相应项为0来求最大值。必须要注意到：$\mathbb{E}^{(w)}_{\gamma_j^{(i)}}$是一个常量，它是根据第$w$轮迭代后的模型结果推算出来的期望值。

- **对α_j求导。** 我们首先对α_j求导数，注意到还有约束$\sum_{j=1}^{k} \alpha_j = 1$，考虑使用Lagrange乘子法来求解，令

  $$L = \sum_{j=1}^{k} \left(\ln \alpha_j \cdot \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} + \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \ln f_j(\boldsymbol{y}^{(i)}) \right) + \lambda \left(\sum_{j=1}^{k} \alpha_j - 1 \right)$$

  于是令

  $$\frac{\partial}{\partial \alpha_j} L = \frac{1}{\alpha_j} \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} + \lambda = 0$$

  得到

  $$\alpha_j = -\frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}}$$

  根据约束条件又有

  $$\sum_{j=1}^{k} \alpha_j = \sum_{j=1}^{k} \left(-\frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \right) = -\frac{1}{\lambda} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{E}^{w}_{\gamma_j^{(i)}} = 1$$

  于是有

  $$\begin{aligned} \lambda &= -\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{E}^{w}_{\gamma_j^{(i)}} \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{k} \frac{\alpha_j^{(i)} f_j^{(w)}(\boldsymbol{y}^{(i)})}{\sum_{t=1}^{k} \alpha_t^{(i)} f_t^{(w)}(\boldsymbol{y}^{(i)})} \\ &= -\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \alpha_j^{(i)} f_j^{(w)}(\boldsymbol{y}^{(i)})}{\sum_{t=1}^{k} \alpha_t^{(i)} f_t^{(w)}(\boldsymbol{y}^{(i)})} \\ &= -n \end{aligned}$$

带入到式(3.2)中，就有

$$\alpha_j = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}} \tag{21}$$

- **对$\boldsymbol{\mu}_j$求导**。注意到式(14)是一个实值函数，因此实值函数对一个向量$\boldsymbol{\mu}_j \in \mathbb{R}^m$求导，结果仍然是一个向量，即实值函数对每一个分量求导。我们先来证明，如果矩阵$\boldsymbol{A}$是一个对称矩阵，则

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) = 2\boldsymbol{A}\boldsymbol{x} \tag{22}$$

事实上，因为

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_i\sum_j a_{ij}x_ix_j$$

对于某个分量x_h，我们可以将上式中与a_h无关的略去，从而得到

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \overset{a_h}{\sim} \sum_{i\neq h} a_{ih}x_ix_h + \sum_{j\neq h} a_{hj}x_jx_h + a_{hh}x_h^2$$

因为$\boldsymbol{A}$是对称矩阵，因此$a_{jh} = a_{hj}$，所以

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x_h}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) \overset{a_h}{\sim}& \frac{\mathrm{d}}{\mathrm{d}x_h}\left[\left(2\sum_{i\neq h}a_{ih}x_ix_h\right) + a_{hh}x_h^2\right]\\ =&2\left(\sum_{i\neq h}a_{ih}x_i\right) + 2a_{hh}x_h\\ =&2\sum_i a_{ih}x_i = 2\sum_i a_{ih}x_i\end{aligned}$$

从而

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) &= \left[\frac{\mathrm{d}}{\mathrm{d}x_1}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) \quad \frac{\mathrm{d}}{\mathrm{d}x_2}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) \quad \cdots \quad \frac{\mathrm{d}}{\mathrm{d}x_m}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right)\right]^T\\ &= \left[2\textstyle\sum_i a_{1i}x_i \quad 2\sum_i a_{2i}x_i \quad \cdots \quad 2\sum_i a_{mi}x_i\right]^T\\ &= 2\boldsymbol{A}\boldsymbol{x}\end{aligned}$$

另外我们还可以推出

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right) &= 2\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\\ \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{s}}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right) &= -2\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\end{aligned}$$

这是因为

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x_i}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right) &= \frac{\mathrm{d}}{\mathrm{d}(x_i-s_i)}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right)\\ \frac{\mathrm{d}}{\mathrm{d}s_i}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right) &= -\frac{\mathrm{d}}{\mathrm{d}(x_i-s_i)}\left((\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{s})\right)\end{aligned}$$

现在我们来求$\boldsymbol{\mu}_j$的导数。注意到式(14)中与$\boldsymbol{\mu}_j$有关的项为

$$Q_j = \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \ln f_j(\boldsymbol{y}^{(i)}) \tag{23}$$

其中

$$\begin{aligned}
&\ln f_j(\boldsymbol{y}^{(i)}) \\
=&\ln \left[\frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j)^T {\boldsymbol{\Sigma}_j}^{-1} (\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j) \right) \right] \\
=&-\frac{1}{2}\left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) + C
\end{aligned}$$

其中$C = -\ln\left((2\pi)^{m/2}|\boldsymbol{\Sigma}_j|^{1/2}\right)$是与$\boldsymbol{\mu}_j$无关的项。注意到$\boldsymbol{\Sigma}_j$是协方差矩阵，因此它是一个对称矩阵，且假设逆矩阵存在，从而对称矩阵的逆矩阵也是一个对称矩阵，于是根据刚才的定理就有

$$\frac{\partial}{\partial \boldsymbol{\mu}_j}\left(\ln f_j(\boldsymbol{y}^{(i)})\right) = \boldsymbol{\Sigma}_j^{-1}\left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) \tag{24}$$

现在我们有

$$\begin{aligned}
\frac{\partial Q_j}{\partial \boldsymbol{\mu}_j} &= \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \frac{\partial}{\partial \boldsymbol{\mu}_j} \ln f_j(\boldsymbol{y}^{(i)}) \\
&= \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \boldsymbol{\Sigma}_j^{-1}\left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) \\
&= \boldsymbol{\Sigma}_j^{-1} \left(\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) \right)
\end{aligned} \tag{25}$$

我们令导向量为$\mathbf{0}$，于是就有

$$\boldsymbol{\Sigma}_j^{-1} \left[\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) \right] = \mathbf{0}$$

因为$\boldsymbol{\Sigma}_j^{-1}$是可逆矩阵，于是就意味着

$$\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right) = \mathbf{0}$$

化简后就得到

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \boldsymbol{y}^{(i)}}{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}}} \tag{26}$$

我们是从整体上来考虑了对均值向量$\boldsymbol{\mu}_j$求导，其关键在于我们将可逆矩阵$\boldsymbol{\Sigma}_j^{-1}$提出来并消去，从而解出$\boldsymbol{\mu}_j$。这就是说，如果按照对分量，

如第h个分量μ_{j_h}，去考虑，单独对它求导，我们可能无法得出想要的结果，具体来说，如果令σ_{ij}是$\mathbf{\Sigma}_J^{-1}$中的元素，那么我们可能会求出

$$\frac{\partial Q_j}{\partial \mu_{j_h}} = \sum_{i=1}^{n}\left(\mathbb{E}^w_{\gamma_j^{(i)}} \sum_{\xi=1}^{m} \sigma_{h,\xi}(y_\xi^{(i)} - \mu_{j_\xi})\right) = \sum_{\xi=1}^{m}\left(\sigma_{h,\xi} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi})\right)$$

令其为0，从中分解出μ_{j_h}，就会得到

$$\mu_{j_h} = \frac{\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}} y_h^{(i)}}{\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}} + \frac{\sum_{\xi\neq h} \sigma_{h,\xi} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j,\xi})}{\sigma_{h,h} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}}$$

会发现其他μ_{j_ξ}无法消去，因此关键点还是在于我们要求出所有分量的导数，并同时令其为0，就有

$$\begin{cases} \frac{\partial Q_j}{\partial \mu_{j_1}} = \sum_{\xi=1}^{m}\left(\sigma_{1,\xi} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi})\right) = 0 \\ \frac{\partial Q_j}{\partial \mu_{j_2}} = \sum_{\xi=1}^{m}\left(\sigma_{2,\xi} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi})\right) = 0 \\ \vdots \\ \frac{\partial Q_j}{\partial \mu_{j_k}} = \sum_{\xi=1}^{m}\left(\sigma_{k,\xi} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi})\right) = 0 \end{cases}$$

如果我们仔细观察这个联立式，就会发现数列$\{\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi})\}, \xi = 1, 2, \ldots, k$构成了可逆矩阵$\mathbf{\Sigma}_j^{-1}$列向量的一个线性组合系数，因为列向量是线性无关的，因此线性组合为0当且仅当$\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}(y_\xi^{(i)} - \mu_{j_\xi}) = 0, \xi = 1, 2, \ldots, k$，从而也能得出

$$\mu_{j_\xi} = \frac{\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}} y_\xi^{(i)}}{\sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}}} \quad \xi = 1, 2, \ldots, k$$

- **对$\mathbf{\Sigma}_j$求导**。从式(14)中删去与$\mathbf{\Sigma}_j$（求导）无关的项，并展开f_j，就得到

$$Q_j = -\frac{1}{2} \sum_{i=1}^{n} \mathbb{E}^w_{\gamma_j^{(i)}} \left(\ln|\mathbf{\Sigma}_j| + \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right)^T \mathbf{\Sigma}_j^{-1} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\right)\right)$$

根据Matrix Cookbook[2]公式(57)和(67)

$$\frac{\partial}{\partial \boldsymbol{X}} \ln|\boldsymbol{X}| = \boldsymbol{X}^{-T} \tag{27}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \left(\boldsymbol{a}^T \boldsymbol{X}^{-1} \boldsymbol{b}\right) = -\boldsymbol{X}^{-T} \boldsymbol{a} \boldsymbol{b}^T \boldsymbol{X}^{-T} \tag{28}$$

[2] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

我们有

$$\begin{aligned}\frac{\partial Q_j}{\partial \boldsymbol{\Sigma}_j} &= -\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}\frac{\partial Q_j}{\partial \boldsymbol{\Sigma}_j}\left(\ln|\boldsymbol{\Sigma}_j| + \left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)^T\boldsymbol{\Sigma}_j^{-1}\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)\right)\\ &= -\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}\left(\boldsymbol{\Sigma}_j^{-T} - \boldsymbol{\Sigma}_j^{-T}\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)^T\boldsymbol{\Sigma}_j^{-T}\right)\\ &= -\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}\boldsymbol{\Sigma}_j^{-T}\left(\boldsymbol{\Sigma}_j - \left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)^T\right)\boldsymbol{\Sigma}_j^{-T}\\ &= \boldsymbol{\Sigma}_j^{-T}\left[-\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}\left(\boldsymbol{\Sigma}_j - \left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)^T\right)\right]\boldsymbol{\Sigma}_j^{-T}\end{aligned} \tag{29}$$

令其为$\boldsymbol{O}$，消去可逆矩阵$\boldsymbol{\Sigma}_j^{-T}$，我们就得到

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)\left(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}_j\right)^T}{\sum_{i=1}^{n}\mathbb{E}^{w}_{\gamma_j^{(i)}}} \tag{30}$$

3.3 总结

根据前面的推导，算法1总结了EM算法在高斯混合模型中的应用。作为一个例子，图4显示了EM算法拟合由$\mathcal{N}(6.5.5)$和$\mathcal{N}(10, 4.3)$两个分布所产生的样本的混合高斯分布的迭代过程。

算法 1 高斯混合模型参数估计的EM算法

输入： 样本集合$Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}$

高斯混合模型（确定混合模型个数k）

输出： 高斯混合模型参数$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Phi$

1: 给定一个初始参数$\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \Phi^{(0)}$,迭代轮次$w = 0$

2: E步：依据当前模型$\boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}$计算极大似然函数对隐变量$\boldsymbol{\gamma}$的期望

$$\mathbb{E}^{w}_{\gamma_j^{(i)}} = \frac{\alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})}{\sum_{j=1}^{k} \alpha_j^{(w)} f_j^{(w)}(\boldsymbol{y}^{(i)})} \quad j = 1, \ldots, k; i = 1, \ldots, n$$

其中

$$f_j^{(w)}(\boldsymbol{y}^{(i)}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_j^{(w)}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)})^T {\boldsymbol{\Sigma}_j^{(w)}}^{-1} (\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)})\right)$$

3: 计算新一轮迭代参数，其中$j = 1, 2, \ldots, k$

$$\alpha_j^{(w+1)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \tag{31}$$

$$\boldsymbol{\mu}_j^{(w+1)} = \frac{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \boldsymbol{y}^{(i)}}{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}}} \tag{32}$$

$$\boldsymbol{\Sigma}_j^{(w+1)} = \frac{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}} \left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)}\right)\left(\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j^{(w)}\right)^T}{\sum_{i=1}^{n} \mathbb{E}^{w}_{\gamma_j^{(i)}}} \tag{33}$$

4: $w \leftarrow w + 1$

5: 若本轮迭代收敛，返回模型参数$\boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(w)}, \Phi^{(w)}$；否则返回第2步。

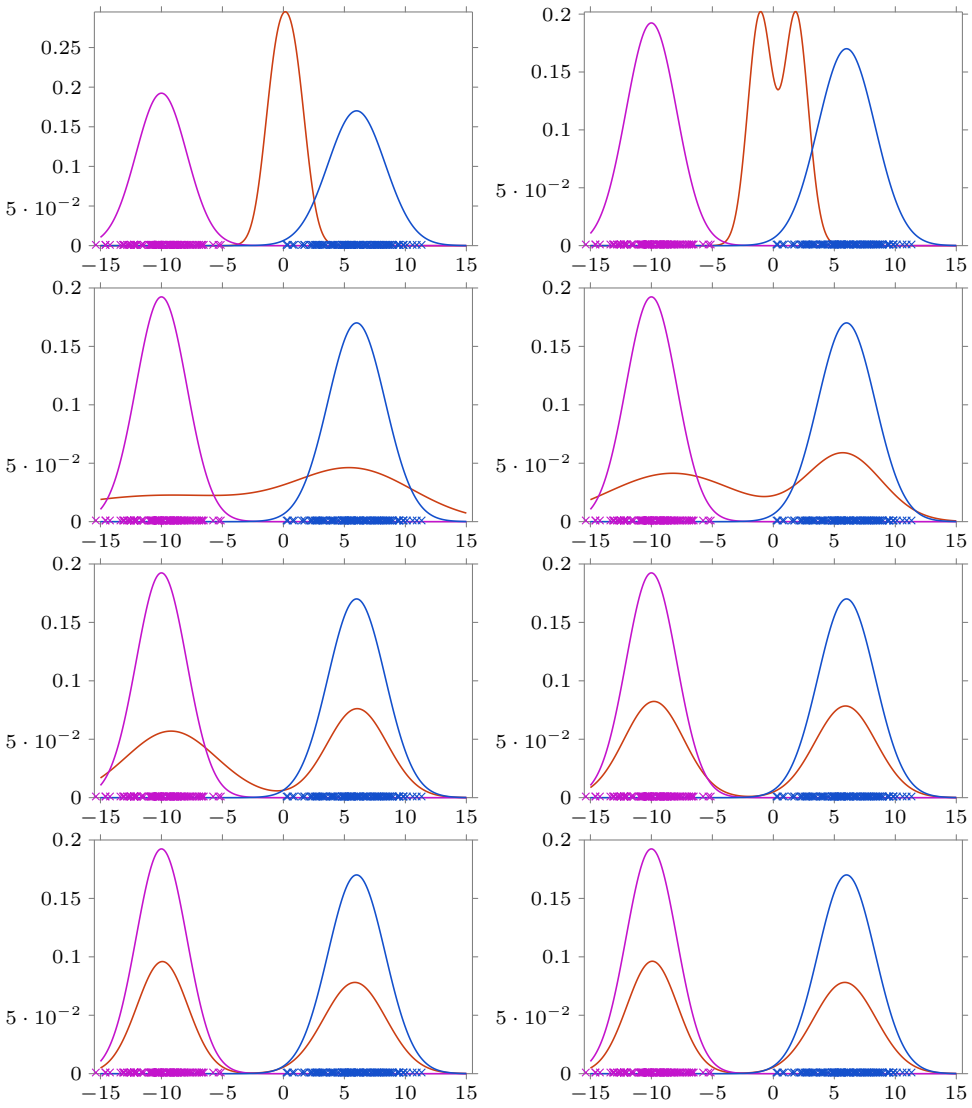

图 4: 一维高斯分布的拟合曲线。其中样本点由$\mathcal{N}(6.5.5)$和$\mathcal{N}(10, 4.3)$两个分布所产生。